DS 501 Case Study 1

Dane Johnson

2/26/2021

Problem 1: Sampling Twitter Data with Streaming API about a certain topic

```
# Collect tweets about a topic from twitter.
setup_twitter_oauth(consumerKey, consumerSecret, accessToken, accessTokenSecret)
```

```
## [1] "Using direct authentication"
```

```
tweets = searchTwitter('#math', n=500)
tweets = strip_retweets(tweets, strip_manual=TRUE, strip_mt=TRUE)
tweetsDF = twListToDF(tweets)

# Store the downloaded tweets into a local file.
write.csv(tweetsDF, file = "tweet_file.csv")
```

- The topic of interest: < Math Related Tweets>
- The total number of tweets collected: < 500 >

Problem 2: Analyzing Tweets and Tweet Entities with Frequency Analysis

1. Word Count:

- Use the tweets you collected in Problem 1, and compute the frequencies of the words being used in these tweets.
- Display a table of the top 30 words (ONLY) with their counts

```
# To analyze word frequency, create a corpus
#from the tweets and then process the corpus
#to remove characters that do not represent
#natural language. Then sort the pared down
#results in decreasing order.
tweets_text <- sapply(tweets, function(x) x$getText())

corpus <- Corpus(VectorSource(tweets_text))

corpus = tm_map(corpus, content_transformer(tolower))
corpus = tm_map(corpus, removeNumbers)
corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, removeWords, stopwords())
corpus = tm_map(corpus, stemDocument)
corpus = tm_map(corpus, stripWhitespace)
corpus = tm_map(corpus, removeWords, c("RT", "are","that", "..."))
removeURL <- function(x) gsub("http[[:alnum:]]*", "", x)
corpus <- tm_map(corpus, content_transformer(removeURL))
```

```
tweets_2 <- TermDocumentMatrix(corpus)
tweets_2 <- as.matrix(tweets_2)
tweets_2 <- sort(rowSums(tweets_2),decreasing=TRUE)
tweets_2 <- data.frame(word = names(tweets_2),freq=tweets_2)

head(tweets_2,30)
```

```
##                     word freq
## math                math  106
## essay              essay   92
## paper              paper   66
## class              class   50
## onlineclass onlineclass   47
## case                case   46
## fall                fall   46
## javascript   javascript   43
## biolog           biolog   41
## busi                busi   39
## studi              studi   38
## calculus       calculus   36
## algebra         algebra   34
## help               help   31
## stat               stat   30
## homeworks...   homeworks...   28
## strain            strain   26
## statist          statist   20
## plan               plan   19
## workout          workout   19
## biometrics... biometrics...   18
## nurs               nurs   17
## need               need   15
## dont               dont   14
## miss               miss   14
## deadlin          deadlin   13
## academ            academ   12
## chemistri     chemistri   12
## anytim            anytim   12
## assistancev assistancev   12
```

2. Find the most popular tweets in your collection of tweets

Please display a table of the top 10 tweets

```
tweetsDF = tweetsDF[order(-tweetsDF$retweetCount),]
most_popular_tweets_DF = data.frame("PopularTweet" = tweetsDF$text,
    "RetweetCount" =tweetsDF$retweetCount)
head(most_popular_tweets_DF, 10)
```

```
##
## 1                Ready? Membership is now Open https://t.co/79BHjlqMAE  #100DaysOfCode #Python #data
## 2          I'm a color-neutral #AntiRacist factorian.\n\nThe history of #math was written by Whites
## 3      Gradient Descent for Machine Learning (ML) 101 with Python Tutorial <U+2192> https://t.co/4dJ
## 4              NeuralNetwork in MachineLearning \n#Bioinformatics #DataVisualization #DataScience \n#
## 5  POPPI <U+2228> PIPOP?\n\n#mathart #gamedev #math #indiegame #indiegamedev #education #indiedev #i
## 6            #DigitalTransformation #MachineLearning #BigData #ArtificialIntelligence #cybersecurity
```

```
## 7          Strain no More in Your,\n#onlineclass\n#fall classes \n#Essay due\n#Paper pay \n#javascri
## 8                              Thinking ahead to the return of our revision groups, I've added a "Re
## 9             DM us for quality grades for your online classes.\n\n#Paperpay\n#python\n#chemistry \n
## 10         Strain no More in Your,\n#onlineclass\n#fall classes \n#Essay due\n#Paper pay \n#javascri
##    RetweetCount
## 1            25
## 2            18
## 3            13
## 4             8
## 5             8
## 6             5
## 7             5
## 8             5
## 9             4
## 10            4
```

3. Find the most popular Tweet Entities in your collection of tweets

Please display a table of the top 10 hashtags (ONLY), top 10 user mentions (ONLY) that are the most popular in your collection of tweets.

```
entity_list = c()
# Create a for statement to populate the list
for (i in seq(1, length(tweets_text), by=1)) {
    entity_list[[i]] = str_extract_all(tweets_text[i], "#\\S+", simplify =  TRUE)
}

entity_list = flatten(entity_list)
entity_corpus = Corpus(VectorSource(entity_list))
entity_corpus = TermDocumentMatrix(entity_corpus)
entity_corpus = as.matrix(entity_corpus)
entity_corpus = sort(rowSums(entity_corpus),decreasing=TRUE)
entity_corpus = data.frame(hashtag = names(entity_corpus),freq=entity_corpus)
head(entity_corpus,10)
```

```
##                   hashtag freq
## #math               #math   91
## #essay             #essay   60
## #paper             #paper   49
## #onlineclass #onlineclass   45
## #fall               #fall   45
## #case               #case   44
## #javascript   #javascript   43
## #biology         #biology   41
## #business       #business   38
## #calculus       #calculus   35
```

```
mention_list = c()
# Create a for statement to populate the list
for (i in seq(1, length(tweets_text), by=1)) {
    mention_list[[i]] = str_extract_all(tweets_text[i], "@\\S+", simplify =  TRUE)
}

mention_list = flatten(mention_list)
mention_corpus = Corpus(VectorSource(mention_list))
mention_corpus = TermDocumentMatrix(mention_corpus)
```

```
mention_corpus = as.matrix(mention_corpus)
mention_corpus = sort(rowSums(mention_corpus),decreasing=TRUE)
mention_corpus = data.frame(user = names(mention_corpus),freq=mention_corpus)
head(mention_corpus,10)
```

```
##                              user freq
## @essayassignmen8 @essayassignmen8    2
## @gmail.com             @gmail.com    2
## @custompapers4     @custompapers4    1
## @shasha11224         @shasha11224    1
## @jamestanton         @jamestanton    1
## @globalmathproj   @globalmathproj    1
## @mathemalicious   @mathemalicious    1
## @lilmathgirl         @lilmathgirl    1
## @mathplay3             @mathplay3    1
## @gamesbygord         @gamesbygord    1
```

Problem 3: Getting any 20 friends and any 20 followers of a popular user in twitter

```
# Twitter User of Interest: Seattle Seahawks Quarterback Russell Wilson

user = getUser("DangeRussWilson")
user$getDescription()
```

```
## [1] "I want to Love like Jesus!"
```

```
#Finding 20 Friends
friends = user$getFriends(n=20)
friendsDF = twListToDF(friends)
friendsDF = data.frame("FriendID" = friendsDF$id, "FriendName" = friendsDF$screenName)
head(friendsDF, 20)
```

```
##               FriendID     FriendName
## 1              5638862          kabir
## 2            582163242   KennyDichter
## 3             50393960      BillGates
## 4             18228898     johnlegend
## 5            161801527   melindagates
## 6             33995409     DwyaneWade
## 7  1238702303244214272 AllHumanNation
## 8             16228398         mcuban
## 9           1366057639          dkm14
## 10           409486555  MichelleObama
## 11            30354991   KamalaHarris
## 12              939091       JoeBiden
## 13              813286    BarackObama
## 14 1019297113560059904     Wnyacademy
## 15            53853197            CP3
## 16           456988174      RandyMoss
## 17            48436234  ApplePodcasts
## 18 1250495274998358016     dangertalk
```

```r
#Finding 20 followers
followers = user$getFollowers(n=20)
followersDF = twListToDF(followers)
followersDF = data.frame("FollowerID" = followersDF$id, "FollowerName" =followersDF$screenName)
head(followersDF, 20)
```

```
##             FollowerID    FollowerName
## 1           2986718321       BTaysty21
## 2   995115035902074880     2003newjohn
## 3  1035650387578109952       BearsfanA
## 4            942805495         bet_bou
## 5           4923381908       ballhog56
## 6   756828860101062656   RealCocailina
## 7  1193318534656536576         wbgulla
## 8  1190666024040980480         Bmix781
## 9  1173487047086170112 Christi06851376
## 10          2914925400    _brandonn_80
## 11 1336397264273764352     Jay95488307
## 12           904058311   evntplnrjulie
## 13           275050636     mannyshow84
```

```r
#Finding users that are both friends and followers
friend_count = user$getFriendsCount()
follower_count = user$getFollowersCount()
all_friends = user$getFriends(n=friend_count)

# The line below will get all followers but exceeds
# Twitter's rate limits since there are so many
# followers. For future projects try using the
# rtweet package instead of twitteR package.

#all_followers = user$getFollowers(n= follower_count)

# So to perform an approximation for this
# assignment, use the same command but
# request some smaller number of followers.

all_followers = user$getFollowers(n=10000)

friends_and_followers=
  intersect(all_friends,all_followers)
friends_and_followersDF =
  twListToDF(friends_and_followers)
friends_and_followersDF =
  data.frame("Friend/FollowerID" = friends_and_followersDF$id,
             "Friend/FollowerName" =
               friends_and_followersDF$screenName)

# However, this still doesn't quite solve the problem.
# Since this twitter account has so many followers,
# even using a large number like 10,000 followers still
# results in a small intersection between this user's
# friends and followers, often just one user.
head(friends_and_followersDF, 10)
```

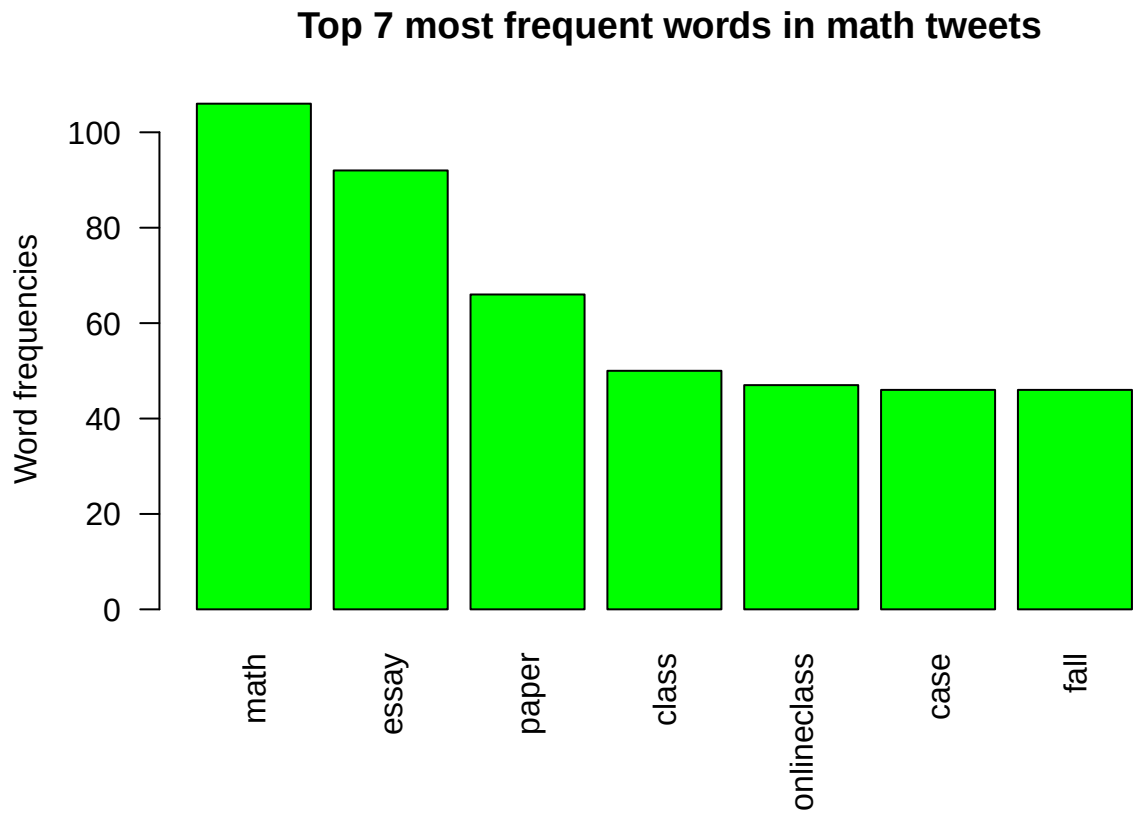
```
##   Friend.FollowerID Friend.FollowerName
## 1        2986718321           BTaysty21
```

Problem 4 (Optional): Explore the data

Run some additional experiments with your data to gain familiarity with the twitter data and twitter API

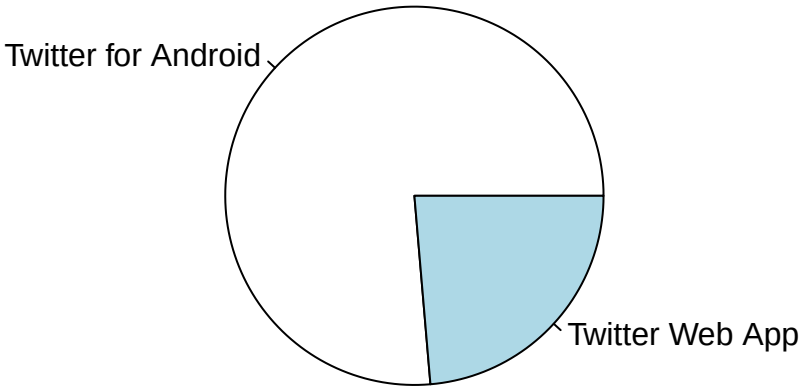
First more exploration with the math tweets.

```
# Most frequent words found in the tweets as barplot
barplot(tweets_2[1:7,]$freq, las = 2,
        names.arg = tweets_2[1:7,]$word,
        col = "green",
        main =" Top 7 most frequent words in math tweets",
        ylab = "Word frequencies")
```

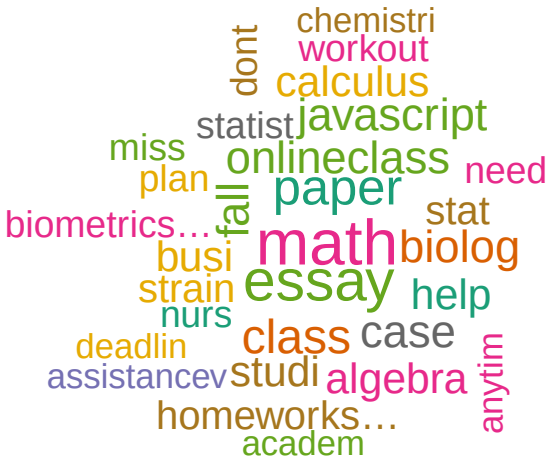



```
# Discover how people are using twitter (what sorts of devices, etc.?)

sources <- sapply(tweets, function(x) x$getStatusSource())
sources <- gsub("</a>", "", sources)
sources <- strsplit(sources, ">")
sources <- sapply(sources, function(x) ifelse(length(x) > 1, x[2], x[1]))
source_table = table(sources)
pie(source_table[source_table > 10])
```




```
# Generate a worldcloud to visualize word frequency.
wordcloud(corpus, min.freq=1,max.words=30,
          scale=c(2,1),
          colors=brewer.pal(8, "Dark2"),
          random.color=T, random.order=F)
```



Next, further exploration with Russell Wilson's Twitter Account

```
total_friends = user$getFriendsCount()
print(total_friends)
```

```
## [1] 91
```

```
recent_tweets = userTimeline("DangeRussWilson")
print(recent_tweets[1:3])
```

```
## [[1]]
## [1] "DangeRussWilson: <U+0001F64C><U+0001F3FE> All for YOUR Glory <U+0001F64F><U+0001F3FE> https://t
##
## [[2]]
## [1] "DangeRussWilson: Lockdown!!!! No one better all year! I see you sis! <U+0001F64F><U+0001F3FE><U
##
## [[3]]
## [1] "DangeRussWilson: All fuel. https://t.co/LvlXeOMpmJ"
```

Final Commentary

For this case study I collected tweets related to math by searching twitter using #math. Originally I collected 500 tweets but stripped this collection of retweets in order to reduce repetition during analysis. This topic is interesting to me since I am a PhD student in math and so I want to know how this field is talked about on social media. I analyzed the data by determining the most frequent words used in tweets about math, what the most common hashtags were in tweets about math, and the most often mentioned twitter users in these tweets. I found that tweets about math are very much dominated by tweets related to homework help for

students. Also, twitter for android is most commonly used to interact with twitter (at least when it comes to the collection I made). An issue I noticed with the tweets collected using searchTweets() is that tweets are frequently truncated so many of the frequent 'words' I found have a letter or a few letters chopped off at the end, i.e. 'biolog' instead of 'biology'.

I also used the twitteR package to learn about how one can study a particular twitter user. I picked Russell Wilson, a professional football player. I am a fan of the Seattle Seahawks.